Explaining Vaccination Rates through Economic Factors

Joshua White

Abstract

At the time of writing, the COVID-19 Pandemic has dominated most people's lives throughout the better part of the last two years. This paper seeks to explain hesitancy in vaccination rates among the 50 US states. To do this, we use both a simple and multiple linear regression model to uncover the relationship between median income and the vaccination rate of a state. In my research, I could not find any previous literature that analyzed this relationship, but the results verify previous research into social factors that explain vaccination rates. A positive correlation between median income and vaccination rates is hypothesized and confirmed by this paper.

I. Introduction

This paper seeks to find a link between vaccination rates and economic variables. The primary variable of interest in this paper is median income, although it may be interesting to analyze GDP/capita as well. The pandemic has been a large part of my life over the past two years; there have been over 250 million cases and 5 million deaths at the time of writing. Additionally, the additional burden of the virus on the healthcare systems, in the United States and worldwide, has undoubtedly resulted in the suffering of many more people.

Currently, in the United States, just over 60% of the population is fully vaccinated. However, this does not show the situation in my home state, Georgia. Here, less than half of the population is vaccinated. Meanwhile, in the neighboring state of Florida, 62% of the population is vaccinated. Theoretically, vaccination would result in the end of the pandemic, but, despite the widely available vaccine, there are many who are skeptical of its efficacy and refuse to get the vaccine. Thus, I believe it is important to look at factors that influence vaccination rates, including factors that result in hesitancy.

I predict that median income is positively correlated with vaccination rates, as it is correlated with education, and I expect more educated individuals to be more likely to obtain a vaccination. Additionally, higher median income generally means that the individuals have a higher standard of living, meaning they have more access to vaccinations (at least in earlier stages of the pandemic, when the vaccine was not as widely available).

II. Literature Review

One lens that the pandemic has been viewed through is that of inequality. Aslan, Marcella, et al. (2021) measure the overall initial health impact of the pandemic. From the beginning of the pandemic until the end of 2020, around 378000 people have died of COVID-19 (Aslan, Marcella, et al. 2021). They discuss challenges in gathering data as well as challenges in quantifying the data, including the possibility that death counts are underreported. They analyze the effects on the pandemic on the Demand and Supply side of healthcare. On the demand side, they note that it is well documented that people have been avoiding medical care because of the pandemic; on the supply side, many nonemergency services were suspended (Aslan, Marcella, et al. 2021). With all of this in mind, they use the metric of excess deaths, which analyzes how far off the deaths in the pandemic period are from a linear trend of rising deaths in non-pandemic years. Using this metric, and controlling for age, they find that Black, Hispanic,

and Native American communities suffered more from the COVID-19 Pandemic than White Americans (Aslan, Marcella, et al. 2021). They also analyze possible reasons why this inequity exists; possible causes are access to COVID-19 information or lack of quality healthcare. The study concludes that Black and Hispanic groups face institutional disadvantages that contributed to their disproportionate suffering in the early stages of the pandemic (Aslan, Marcella, et al. 2021).

Another aspect of the pandemic is the vaccine rollout. In December 2020, The FDA granted emergency use access for two vaccines against COVID-19 (Larson, et al. 2021). As part of a report on Vaccine Confidence, Larson, Morrison et al. discuss the rollout of the vaccine, inequities present in that rollout, and possible reasons why individuals in the US may be vaccine hesitant. They find that data shows in the early stages of the rollout, whites obtained a disproportionately high percentage of vaccines, which may have been accounted for in differential access to registration services and availability to wait at vaccination sites (Larson, et al. 2021). The report discusses a brief history of vaccine hesitancy, noting that hesitancy is "context specific" (Larson, et al. 2021), and provides many different reasons people may have for being hesitant, including "Safety Concerns", "A Legacy of Discrimination", "Political Arguments", "Conspiracy Theories", "Alternative Health", "Government Requirements", and "Religious or Moral Objections" (Larson, et al. 2021). Of particular interest in this paper are the "Legacy of Discrimination" and "Political Arguments". Historically disadvantaged communities have been hit the hardest by the Covid-19 pandemic, and there is a legacy of discrimination and government abuse against them, so, their questions about the vaccine are not rooted in misunderstanding, but in mistrust of institutions (Larson, et al. 2021). Additionally, the pandemic has exposed differing levels of hesitancy on political lines, as well as the urban/rural divide (Larson, et al. 2021).

Yet another perspective on the pandemic is that of the Labor Market. It is no secret that in the early stages of the pandemic, the virus forced people into unemployment by the beginning of the second half of 2020 (Gallant, et al. 2020). Gallant, et al. developed a model to analyze the unique nature of the initial shock of the pandemic. In particular, the model distinguishes between "temporary unemployment", individuals who expect to be recalled by their formal employer, and "permanent unemployment", those who are permanently separated (Gallant, et al. 2020). There is also a distinction drawn between the temporary unemployed who search for jobs while they wait to be recalled, and those who do not, as those who search for jobs in their wait contribute to "congestion" in the labor market (Gallant, et al. 2020). As they note, 24% of

workers say that jobs are hard to find, even though unemployment is higher than at any point during the Great Recession (Gallant, et al. 2020). This model may help understand this fact; temporarily unemployed people may still be considered unemployed by the government.

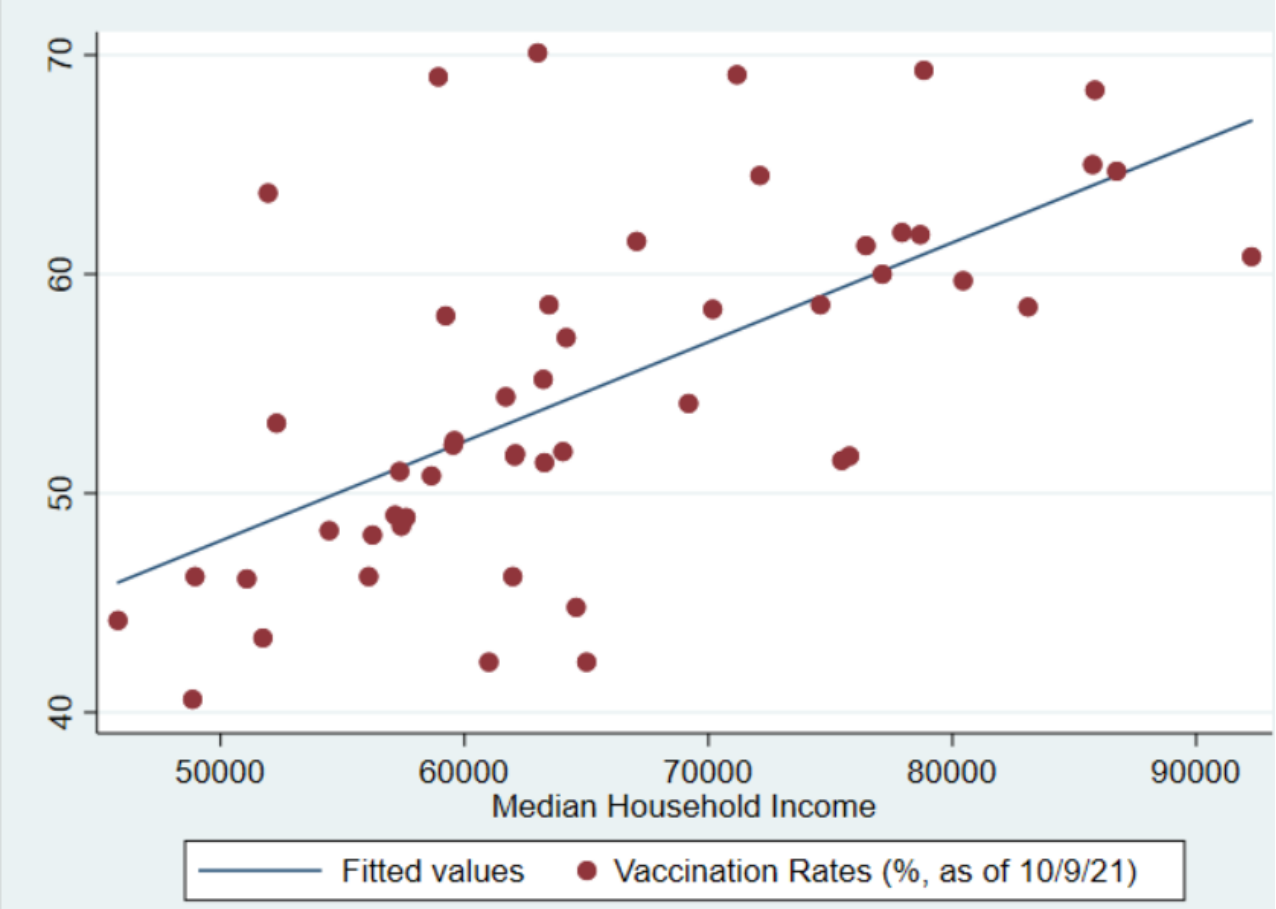
Although there is research into vaccines from an economics perspective, I could not find any literature specifically on analyzing vaccine rates as a function of economic indicators. Much research on the COVID-19 Pandemic has been focused on the inequalities exposed by the pandemic, such as minority groups in the United States (Aslan, Marcella, et. al. 2021), of vaccine rollout (Felter, 2020), and of the labor market (Gallant, et al. 2020), but nothing on the link between economic factors and vaccinations. Of course, the vaccine has been politicized; former President Donald Trump discouraged his followers from getting the vaccine and generally downplaying the pandemic, but, controlling for this, it may be possible to explain vaccination rates through purely economic factors. Thus, this paper seeks explain disparities in vaccination rates, as well as to pave way for more research to be done into this topic

III. Data

To understand the relationship between median income and vaccination rates, cross-sectional data from the fifty states was gathered. Obviously, pandemics are most virulent in population centers, so it is of interest to control for population, as well as population density. All the sources included data on Washington DC, but DC is a very large outlier, both in terms of population density and GDP/Capita. Thus, I discarded the data obtained from Washington DC. The data on vaccination rates, the variable of interest, is being updated daily. As this paper is interested in cross-sectional data, I am specifically looking at vaccination rates as of October 9, 2021 (as this is when I gathered the data from the CDC website).

Graph 1 – Vaccination Rates vs. Median Income

.



As we can see from these scatterplots, Vaccination rates are positively correlated with Median Income. However, there is a large amount of variance that is unable to be explained by this simple linear regression model. Thus, it only makes sense to control for other variables. From the literature, race plays an impact on vaccine hesitancy (Larson et al. 2021), so I will consider the proportion of the population that is white in my analysis. We are also interested in controlling for the effect of political lean on vaccine rates, as that is another important factor that explains vaccine hesitancy (Larson et al. 2021); however, this will be left as an extension in section 5.

A table summarizing the data can be found below. Additionally, summary statistics on the important variables can be found in the next table.

Table 1

Variable Name	Description	Year	Units	Source
vaxx	Vaccination Rate	2021	Percent of Population (%)	CDC (as of 10/9/2021)
medinc	Median Household Income	2020	United States Dollars	US Census 2020
pop	Population	2020	Number of People	US Census 2020
popwhite	Population that is White	2020	Number of People	US Census 2020
percwhite	Percent of Pop that is White	2020	Percent (%) of Population	Derived
landarea	Land Area	2020	Square Miles (mi^2)	US Census 2020
popdens	Population Density	2020	Number of People per square mile	Derived

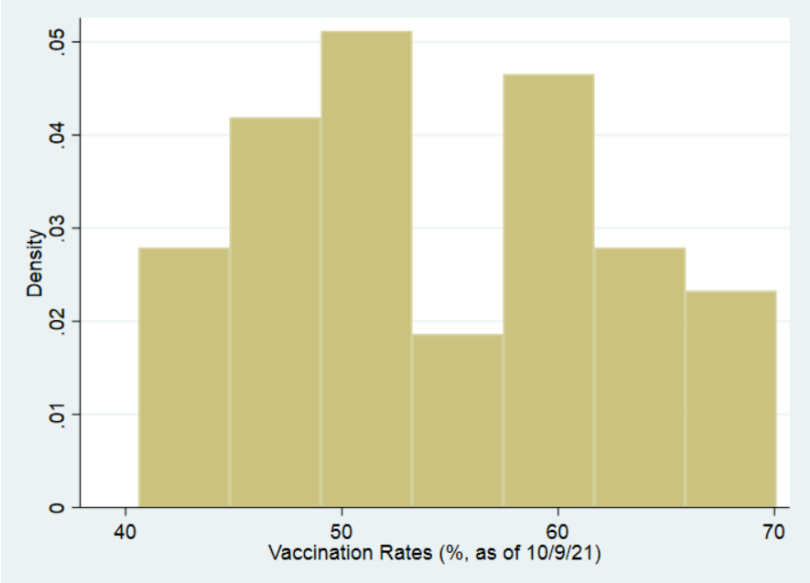
Table 2 - Descriptive Statistics of Important Variables

Variable	#Observations	Mean	Std. Dev.	Min	Max
vaxx	50	54.754	8.04	40.6	70.1
medinc	50	64976.22	10604	45792	86738
percwhite	50	68.521	14.13	22.9	90.801

popdens	50	173.945	214.75	1.1	1064

The number of observations remained consistent as every state has an observation. Additionally, there is a wide variation in the population density of the 50 states. The state with a population density of 1.1 is Alaska, and the state with the highest is New Jersey. We now must check the Classical Linear Model assumptions to determine the usefulness of this model

1. **Linearity:** The model is linear in parameters: We assume that the true model is $y = B_0 + B_1x_1 + \ldots + B_n x_n + u$, where B are the slope coefficients and u is an unobserved error term. All the models created in this paper will be linear in parameters as well, so this assumption is satisfied.
2. **Random:** Essentially all the data was gathered from the 2020 Census (save the vaccination rates, gathered on a particular day in 2021). The Census data was gathered by random sampling, as the Census gathers data through surveys of randomly selected households; however, the CDC data is from the entire population, as they know exactly how many doses have been distributed, and to whom. This means that this data is not random, but this should be fine, as the entire population is considered.
3. **Multicollinearity:** To satisfy this assumption, none of the variables must be perfectly correlated. Using a Stata command, we can check the variables used. As Appendix B shows, there is no perfect collinearity between any of the explanatory variables used. Thus, the assumption holds.
4. **Exogeneity:** We assume that there is no conditional mean. For this to be satisfied, the expected value of the unobserved term must be zero for any possible values of **x.** This may not be satisfied, as there is the possibility of omitted variable bias, or that there is a variable that impacts vaccine hesitancy that this paper does not consider. For now, we will assume this is not the case.
5. **Homoskedasticity:** The variance of the error term is constant. If this were satisfied, the distribution of the y variable will have constant variance. We will check this with assumption 6, so for now, assume this is satisfied.
6. **Normality of Error Distribution:** The population distribution of the error term is normally distributed. Assuming this will imply that the distribution of vaccine rates is also normal. To check this (and assumption 5), here is a distribution of Vaccine Rates:



From this, we see that the data is not quite normal, instead, it looks roughly bimodal. This may have to do with the bucket size of the data. For now, we will assume that this is roughly normal, and thus, has a constant variance. Hence, we assume that the error term is normally distributed with constant variance.

With these assumptions, we now move to creation of the models.

IV. Results

First, we will analyze the simple regression model. This model analyzes the vaccination rates as a function of median income.

Model 1: $vaxx = B_0 + B_1(medinc) + u$

The stata regression results can be found in the appendix. The equation generated from these results is:

Estimated Equation 1: $\widehat{vaxx} = 23.2928 + .0004842(medinc)$

As the minimum value of median income is ~45000, it is important to note that the B1 is small as the magnitude of medinc is in the thousands. This model has an R^2 of .4069, meaning that 40% of the variation in vaccination rates is explained by the model. The correlation coefficient (obtained by taking the square root of R^2 is 0.6425, meaning we have a somewhat strong positive correlation. The intercept is 23.29, but the interpretation of this is not useful as it does not make sense for a state to have a median income of 0. The coefficient B1 is positive, so

there is a positive linear relationship of median income on vaccination rates. As this is a level-level model, the value of B1 means that for every $1 increase in median income, we expect the vaccination rate to increase by .0004842% (as that is the units of vaccination rate). For a more usable interpretation, if median income increases by $1000, we expect the vaccination rate to increase by around .4%. This model is certainly useful, and it does confirm the hypothesis, but, as there is possible omitted variable bias, we should consider a multiple linear regression model to find the ceteris paribus effect of median income on vaccination rates.

Now we will look at the first of our multiple linear regression models:

Model 2: vaxx = = $B_0 + B_1$(medinc) + B_2(popdens) + B_3(percwhite) + u

Again, the stata regression can be found in the appendix. The generated equation is as follows:

Estimated Equation 2: $\widehat{vaxx}$ = 28.953 + .0003902(medinc) + .011 (popdens) + .068(percwhite)

The first thing of note is that the coefficient of percwhite is has a relatively high standard error. In fact, based off the t value, it is not significant at even the 10% significance level. Thus, in future models, it will be disregarded. With the addition of two other explanatory variables, we now have an adjusted R^2 of .4299, which is slightly better than the .4 of the simple linear regression model. Again, the intercept coefficient has no reasonable interpretation, as median income is again never zero. Holding all else constant, the slope coefficient on medinc means that for every $1000 increase in median income, we expect vaccination rates to rise by .3716%. Holding median income and percwhite constant, an increase in population density by 1 person/mi^2 increases the expected vaccination rate by .0107%. This does confirm the idea that a higher population density that was impacted more by the pandemic will have a higher vaccination rate.

With this, we move on to Model 3. In model 3, we drop percwhite due to it not being statistically significant:

Model 3: vaxx = = $B_0 + B_1$(medinc) + B_2(popdens) + u

Estimated Equation 3: $\widehat{vaxx}$ = 28.77 + .0003715(medinc) + .0105(popdens)

There is not a whole lot to say about this model that wasn't said in Model 2. Dropping percwhite reduced the variance of the other terms, which is expected, as dropping irrelevant variables reduces the standard error of the other terms. Furthermore, the Adjusted R^2 value is now .4414, which is slightly higher than model 2. Again, the coefficient on the intercept has no practical interpretation, and the slope coefficients are about the same as they were before.

We now create a table to summarize the results of these models:

Table 3: Regression Model Summary

Independent Variables	**Model 1 (SLR)**	**Model 2 (MLR)**	**Model 3 (MLR)**
medinc	.00048*** (.000084)	.00039*** (.000096)	.00037*** (.000095)
popdens	-	.011** (.0047)	.01** (.0047)
percwhite	-	.068 (.064)	-
intercept	23.29***	22.7***	28.77***
no. observations	50	50	50
Adjusted R^2	.3945	.4429	.4414

*Significant at 10%, **5%, ***1%

Again, the slope coefficients on medinc are all very low, but this is an issue of order of magnitude. This coefficient is always significant in every model, even at the 1% level.

V. Extensions

From reading the literature, political lean plays a part in vaccine hesitancy. As a result, this model should take that into consideration. There are a couple ways that made sense to measure political lean, such as considering presidential election victory margins in various years, but I went with the simplest one, taking in to account the governor of the state.

Variable Name	Description	Year	Units	Source
pollean	1 if Dem Gov. 0 otherwise	2021 (as of 9/16/21)	Binary variable	I don't know what to tell you here, I just googled it.

Variable	Obs	Mean	Std. Dev.	Min	Max
pollean	50	.46	.5034574	0	1

Instead of going through the entire process of discussing the extension models individually, they have both been included in the following table.

Table 4: Extension Regression Summary

Independent Variables	Model (1)	Model (2)	Model (3)	Extension Model (1)	Extension Model (2)
medinc	0.00048*** (.000084)	.00039*** (.000096)	.00037*** (.000095)	.0003426*** (.000088)	.0003196*** (.000089)
popdens	-	.011** (.0047)	.010** (.0047)	.010** (.0043)	.0091** (.0043)
percwhite	-	.068 (.064)	-	.104* (.059)	-

pollean	-	-	-	5.49*** (1.64)	4.95*** (1.65)
intercept	23.29*** (5.55)	22.76*** (8.14)	28.77*** (5.86)	21.00*** (7.39)	30.11*** (5.44)
No. of obs.	50	50	50	50	50
R-square	.4069	.4770	.4642	.5807	.5514
Adj R-square	.3945	.4429	.4414	.5435	.5221

From the table, we can see that political lean is significant at 1% in both models it appears in. Interestingly, the addition of political lean makes percwhite significant at the 10% level.

The two “least significant” variables were percwhite and popdens. To see if these variables are jointly significant, we conduct an F test on the unrestricted model Extension Model 1, and the restricted model that appears in the appendix:

H_0: $B_2 = B_3 = 0$

H_A: H_0 is false

At the 5% significance level, $F_{2, 45} = 3.23$ is our critical value. Our test statistic is F = [(.5807-.5087)/2]/ [(1-.5807)/45] = .036/.0093 = 3.8635. As F > 3.23, we have sufficient evidence to reject the null hypothesis in favor of the alternative, thus, percwhite and popdens are jointly significant at the 5% level.

VI. Conclusions

The hypothesis of a positive correlation between median income and vaccination rates was ultimately confirmed by the regression models. In each model, there was a strong positive correlation between median income and vaccination rates, which did not get weaker with the addition of more explanatory variables.

Most of the secondary explanatory variables also ended up being significant, but percwhite was only significant once political lean was considered. A possible explanation for this is that, on average, people who are white tend to lean Republican, and Republicans are more likely to decline to get vaccinated. Once political lean is considered, the effect of "whiteness" on the vaccination rate is more pronounced. Every other explanatory variable was significant. Of course, in this study, political lean was simply a binary variable.

If I were to do this study again, I would want to look at individual counties instead of the state level. I feel county level data might reveal an even stronger relationship between median income and vaccination rates. Consequentially, considering the political lean of a county is not as simple as the party the governor is aligned with. To fix this, I would change pollean to be a continuous variable more representative of the county's beliefs than the state's. A possible example of this would be the margin of victory in Presidential elections (positive if winner won the county, negative otherwise). Both considerations would likely make a more robust model and better reveal the relationship between median income and vaccination rates.

References

Alsan, Marcella, et al. “The Great Unequalizer: Initial Health Effects of COVID-19 in the United States.” *The Journal of Economic Perspectives*, vol. 35, no. 3, American Economic Association, 2021, pp. 25–46, https://www.jstor.org/stable/27041213.

Larson, Heidi J., et al. “A Crisis Decades in the Making: Vaccine Hesitancy from Smallpox to Covid-19.” *Why Vaccine Confidence Matters to National Security*, Center for Strategic and International Studies (CSIS), 2021, pp. 6–11, http://www.jstor.org/stable/resrep32133.6.

GALLANT, JESSICA, et al. “Temporary Unemployment and Labor Market Dynamics during the COVID-19 Recession.” *Brookings Papers on Economic Activity*, Brookings Institution Press, 2020, pp. 167–216, https://www.jstor.org/stable/27059301.

www.2020census.gov

www.CDC.gov

Appendix A: List of States Used:

Alabama	Montana
Alaska	Nebraska
Arizona	Nevada
Arkansas	New Hampshire
California	New Jersey
Colorado	New Mexico
Connecticut	New York
Delaware	North Carolina
Florida	North Dakota
Georgia	Ohio
Hawaii	Oklahoma
Idaho	Oregon
Illinois	Pennsylvania
Indiana	Rhode Island
Iowa	South Carolina
Kansas	South Dakota
Kentucky	Tennessee
Louisiana	Texas
Maine	Utah
Maryland	Vermont
Massachusetts	Virginia
Michigan	Washington
Minnesota	West Virginia
Mississippi	Wisconsin
Missouri	Wyoming

Appendix B: Regression Models

Collinearity

```
. correlate medinc popdens percwhite pollean
(obs=50)
```

	medinc	popdens	percwh~e	pollean
medinc	1.0000			
popdens	0.5269	1.0000		
percwhite	-0.2930	-0.2748	1.0000	
pollean	0.2844	0.2366	-0.2660	1.0000

Simple Regression Model:

```
. regress vaxx medinc
```

Source	SS	df	MS	Number of obs	=	50
				F(1, 48)	=	32.93
Model	1291.74651	1	1291.74651	Prob > F	=	0.0000
Residual	1882.95769	48	39.2282853	R-squared	=	0.4069
				Adj R-squared	=	0.3945
Total	3174.7042	49	64.7898816	Root MSE	=	6.2632

vaxx	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medinc	.0004842	.0000844	5.74	0.000	.0003145	.0006538
_cons	23.29286	5.553677	4.19	0.000	12.12644	34.45927

Multiple Linear Regression Model 2:

```
. regress vaxx medinc popdens percwhite
```

Source	SS	df	MS	Number of obs	=	50
				F(3, 46)	=	13.99
Model	1514.44239	3	504.814131	Prob > F	=	0.0000
Residual	1660.26181	46	36.092648	R-squared	=	0.4770
				Adj R-squared	=	0.4429
Total	3174.7042	49	64.7898816	Root MSE	=	6.0077

vaxx	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medinc	.0003902	.0000968	4.03	0.000	.0001953	.0005851
popdens	.0113065	.0047544	2.38	0.022	.0017365	.0208765
percwhite	.06817	.0642013	1.06	0.294	-.0610605	.1974004
_cons	22.76397	8.149085	2.79	0.008	6.360712	39.16723

MLR Model 3:

```
. regress vaxx medinc popdens
```

Source	SS	df	MS	Number of obs	=	50
				F(2, 47)	=	20.36
Model	1473.74958	2	736.87479	Prob > F	=	0.0000
Residual	1700.95462	47	36.1905238	R-squared	=	0.4642
				Adj R-squared	=	0.4414
Total	3174.7042	49	64.7898816	Root MSE	=	6.0159

vaxx	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medinc	.0003715	.0000954	3.90	0.000	.0001797	.0005634
popdens	.0105584	.0047082	2.24	0.030	.0010867	.0200302
_cons	28.77671	5.868108	4.90	0.000	16.9716	40.58182

Extension Model 1: Political Lean

```
. regress vaxx medinc popdens percwhite pollean
```

Source	SS	df	MS	Number of obs	=	50
				F(4, 45)	=	15.58
Model	1843.6166	4	460.904149	Prob > F	=	0.0000
Residual	1331.0876	45	29.5797245	R-squared	=	0.5807
				Adj R-squared	=	0.5435
Total	3174.7042	49	64.7898816	Root MSE	=	5.4387

vaxx	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medinc	.0003426	.0000888	3.86	0.000	.0001637	.0005214
popdens	.010162	.0043177	2.35	0.023	.0014657	.0188584
percwhite	.1049908	.0591596	1.77	0.083	-.0141626	.2241443
pollean	5.497117	1.647856	3.34	0.002	2.178166	8.816069
_cons	21.00576	7.396088	2.84	0.007	6.109279	35.90225

Extension Model 2:

```
. regress vax medinc popdens pollean
```

Source	SS	df	MS	Number of obs	=	50
				F(3, 46)	=	18.85
Model	1750.45285	3	583.484285	Prob > F	=	0.0000
Residual	1424.25135	46	30.9619858	R-squared	=	0.5514
				Adj R-squared	=	0.5221
Total	3174.7042	49	64.7898816	Root MSE	=	5.5643

vaxx	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medinc	.0003196	.0000899	3.55	0.001	.0001386	.0005005
popdens	.0091636	.0043798	2.09	0.042	.0003475	.0179797
pollean	4.951487	1.656315	2.99	0.004	1.617498	8.285476
_cons	30.11834	5.446212	5.53	0.000	19.15568	41.08099

Extension Model 3 (For F Test):

```
. regress vaxx medinc pollean
```

Source	SS	df	MS	Number of obs	=	50
				F(2, 47)	=	24.33
Model	1614.91727	2	807.458633	Prob > F	=	0.0000
Residual	1559.78693	47	33.1869561	R-squared	=	0.5087
				Adj R-squared	=	0.4878
Total	3174.7042	49	64.7898816	Root MSE	=	5.7608

vaxx	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medinc	.0004124	.000081	5.09	0.000	.0002495	.0005752
pollean	5.320666	1.705036	3.12	0.003	1.890575	8.750757
_cons	25.51299	5.157469	4.95	0.000	15.1375	35.88848